

### **Amendments to the Specification:**

Please amend the paragraph beginning on page 11, line 11, of the specification with the following amended paragraph:

*Step 02.* Substitutions to a sequence of step 01 are identified using a combination of changes to the antibody sequence. Such changes are either in monomer identity or in monomer physico-chemical properties. These changes span either the CDR and/or the framework region of heavy chain and/or the light chain of the antibody[[ ]]. For example, consider the case in which the heavy chain of the antibody is being humanized. In step 02, a determination can be made that the 21<sup>st</sup> and 49<sup>th</sup> positions of the heavy chain (based on the kabat numbering scheme) can be changed. Moreover, in some embodiments, a determination is made as to which substitutions can be made at such positions in step 02. For instance, step 02 may not only determine that the 21<sup>st</sup> position of the antibody can be changed, but may also determine that this position should be changed to a glycine, alanine, or leucine.

Please amend the paragraph beginning on page 11, line 23, of the specification with the following amended paragraph:

In typical embodiments, several independent rules are used to determine which positions of the ~~antibodies~~ antibodies of step 01 can be changed. Each such rule scores or ranks individual substitutions based on different methods and based on the nature of optimization (i.e) humanization or maturation. Representative rules include, but are not limited to, rules based on (i) changes found in functional, structural or sequence classes, (ii) changes predicted to be favorable using substitution matrices, (iii) changes predicted using evolutionary analysis of the antibody structural and sequence classes, (iv) changes seen in random ~~mutagenesis~~ mutagenesis screening, (v) changes predicted by structural modeling, (vi) changes proposed by an expert on the antibody and (vii) changes predicted to be favorable using structural information (vii) changes derived from comparing the framework region of the antibodies with human germline sequences (viii) changes derived from comparing the framework regions of human antibodies (ix) changes derived from substitution matrices constructed from the positional frequencies of

amino acids in the CDR regions of all antibodies. ~~[(x)]~~ Any number of rules can be applied to the one or more antibodies of step 01.

Please amend the paragraph beginning on page 12, line 17, of the specification with the following amended paragraph:

To illustrate, consider the case in which the antibody of step 01 ~~[(,)]~~ is a murine antibody and the 2<sup>nd</sup>, 5<sup>th</sup>, and 15<sup>th</sup> kabat positions of the heavy chain ~~has~~ have been identified as candidate substitution positions in step 03. Assuming that each of these three positions can be independently substituted with any of the twenty naturally occurring amino acids, there are  $20^3 - 1$  different variant antibodies that could be constructed. In some instances, step 02 will constrain the types of amino acids that can be substituted at these positions based on the rules described above. Nevertheless, the full antibody sequence space proposed in step 02 even after filtering can be large. Step 03 seeks to minimize the number of variants that are constructed in order to evenly search and sample this large sequence space.

Please amend the paragraph beginning on page 13, line 21, of the specification with the following amended paragraph:

*Step 08.* The performance of the methods used to select substitution positions in step 02 and to model the sequence-activity relationships in instances of step 05 are assessed by analyzing the sequences of the best performing variants. In general, the best performing variants are any variants in any iteration of the cycle defined by steps 04-07 that score best in one or more functional assays for the target antibody. Step 08 provides a method for tuning the adjustable parameters of the system. Once these parameters have been adjusted, steps 02 through 07, including multiple iterations of the cycle defined by steps 04-07, are repeated. Advantageously, one of the adjustable parameters of the system is the individual weights for each of the methods applied in step 02. For example, those step 02 ~~method~~ methods that were good at identifying substitution positions associated with high scoring antibody variants are up-weighted in the next instance of steps 02 through 07. The modification of weights applied to methods in step 02 based on the results of cycles of steps 04-07 allows the system to learn

from previous results thereby improving the accuracy with which the system can identify beneficial substitutions (in step 02) and assess the contribution of substitutions to antibody activity (in steps 05 and 06).

Please amend the paragraph beginning on page 20, line 31, of the specification with the following amended paragraph:

Inference engine 106 can also use structural information 128 (*e.g.*, crystal ~~structure, *insilico*~~ structure, *in silico* models of antibodies, *de novo* modeled antibody, *etc.*) stored in knowledge base 108. For example, inference engine 106 can assign higher probabilities to amino acid residues in framework regions that are close to the CDR of an antibody, as will affect activity and/or specificity than more distant residues. Similarly, proximity to an epitope, proximity to an area of structural conflict, proximity to a conserved sequence, proximity to a binding site, proximity to a cleft in the protein, proximity to a modification site, *etc.* can be calculated from structural information 128 and used to calculate the probability that a substitution will result in a functional antibody. To calculate the distance of a residue from a region of functional interest, physical distances obtained using a known crystal structure of the reference sequence can be used. Alternatively, molecular modeling approaches can be used. For example, the structure of the reference sequence can be predicted based on its homology to a known structure, and then used to calculate distances. Or the entire structure of the reference sequence can be predicted and distances then calculated from the predicted structure.

Please amend the paragraph beginning on page 21, line 19, of the specification with the following amended paragraph:

In addition to the examples of elements of information that can be used as a part of a knowledge base 108 described above, other information that can contribute to an antibody knowledge base 108 that can then be used by inference engine 106 of an expert system 100 to calculate the probability that a substitution will possess a desired function include, but ~~[[are]]~~ is not limited to, individual sequence analysis (including sequence complexity, sequence content

and composition, internal base-pairing and secondary structure predictions) sequence comparisons (including structure-based sequence alignments, homology-based sequence alignments, phylogenetic comparisons based on multiple pairwise comparisons, phylogenetic comparisons based on principal component analysis of sequence alignments, Hidden Markov models), evolutionary molecular analysis, structural analysis (including those using X-ray crystallographic data, nuclear magnetic resonance studies, structure threading algorithms, molecular dynamic simulations, active site geometry, determination of surface, internal and active site residues), known or predicted data relating sequence or structure to functional mechanisms, chemical and biophysical properties of functional groups, known or predicted functional effects of changes (for example information derived from the Protein Mutant Resource database, from an evolutionary comparison of sequence and activity data or from a comparison binding pockets and ~~residues~~ residues for the antibody with binding pockets and ~~residues~~ residues for other antibodies or sets of antibodies), substitution matrices derived from sequence comparisons, mutations that are known or that can be predicted to affect physical properties of proteins (including stability, thermostability), known or predicted properties (including plasticity and tolerance to substitutions) of homologous or related antibodies (including other members of sequence, structurally or functionally related classes of antibodies), known or predicted immunological effects and constraints for specific sequence residues or motifs, known or predicted sequence effects on *in vivo* or *in vitro* post-translational or post-transcriptional modifications, known or predicted effects of the functional environment (including other proteins, nucleic acids or other molecules contained within a cell), measured or predicted biochemical or biophysical properties (including crystallization), effects of sequences on the expression of nucleic acids or proteins (including known or predicted RNA splice sites, protein splice sites, promoter sequences, transcriptional enhancer sequences, transcription and translation terminator sequences, sequences that affect the stability of a protein or nucleic acid, codon usage tables, nucleic acid GC content). Sources of this information can include, without limitation, text mined from scientific literature, data mined from genomic sequences, expressed sequences, structural databases and in second and subsequent iterations of the process, case specific data from the first points of the sequence space mapped.

Please amend the paragraph beginning on page 24, line 23, of the specification with the following amended paragraph:

There are many variations of ways to combine scores produced by two or more ~~rules 120~~ rules 120. Variations are possible (i) in the methods of assigning scores, (ii) in the methods of combining scores, and (iii) in the methods of assigning different weights to scores produced by different rules 120. Rules 120 can also be combined on a case by case basis, using expert knowledge. These rules 120 can be stored in a knowledge base 108 and can be executed by inference engine 106 using user input acquired by questioning the user for requirements and knowledge via the user ~~interface 104~~ interface 104.

Please amend the paragraph beginning on page 27, line 32, of the specification with the following amended paragraph:

In addition, prior to combination, scores produced by individual rules can be scaled or normalized and/or ~~transformed~~ transformed by a mathematical function to facilitate their combination. For example, in the case humanization of RSV-19, the mutation I46V was identified as the most favorable substitution in the framework region by combining the scores from methods 130. The distance, expressed in fraction of amino acid differences, was transformed and a Poisson correction ( $-\log[1-\text{fraction}]$ ) applied and multiplied by the product of the absolute scores obtained from the other methods 130. The resulting scores for all substitutions were ranked and I46V (combination score 126) was ranked 1. In this example different criteria were used to compute the scores for the framework and CDR regions.

Please amend the paragraph beginning on page 28, line 4, of the specification with the following amended paragraph:

As indicated in Section 5.1.2, the scores produced by individual rules 120 can be assigned different weights prior to being combined. For example, if the total score for a substituting monomer  $x$  at position  $i$  ( $S_{ix}$ ) is obtained by adding the scores obtained by applying  $n$  different rules, the score can be expressed by Equations (1) or (2):

$$(Eq. 1) S_{ix} = W_1 i_x R_1 + W_2 i_x R_2 + W_3 i_x R_3 + W_4 i_x R_4 + W_5 i_x R_5 + \dots + W_n i_x R_n$$

where,

$i_x R_n$  is the score given by rule  $n$  for substituting monomer  $x$  at position  $i$ ; and

$W_n$  is a weight applied to the score given by rule  $n$

$$(Eq. 2) S_{ix} = f(W_1 R_1(i_x), W_2 R_2(i_x), \dots, W_j R_j(i_x))$$

where,

$R_j(i_x)$  is the score given by rule  $j$  for substituting monomer  $x$  at position  $i$ ;

$W_j$  is a weight applied to scores given by rule  $j$ ; and

$f$  is some mathematical function.

Please amend the paragraph beginning on page 31, line 12, of the specification with the following amended paragraph:

One source of information that can be used to construct rules 120 that assess the likely effect of amino acid substitutions upon one or more activities of an antibody is the sequence of one or more homologous or related antibodies. See, for example, Fig. 3, rule 3a. Homologous sequences are generally analogous functionally and structurally, although having been subjected separately to different selective pressures they are also likely to be optimized differently. Antibody sequences variants can also be generated in the lab using many techniques and sequence[[,]] properties of several such antibodies are available in the database and literature. Amino acids that differ between homologous sequences thus provide a guide to substitutions that are likely to yield functional though different antibody sequences. For humanization of antibodies, alignment of the [[the]] target antibody with human germline sequences available in the databases is used to identify ~~residue~~ residues in the human framework. The sequences can be grouped into classes as defined by Chothia and Lesk (Chothia C, Lesk AM, "Canonical structures for the hypervariable regions of immunoglobulins." J Mol Biol. 1987 Aug 20;196(4):901-17). Alignment of homologous sequences can therefore be used to identify candidate substitution positions.

Please amend the paragraph beginning on page 32, line 23, of the specification with the following amended paragraph:

Amino acid diversity and tolerance at each site can be measured as a fitness property of each amino acid at every location. In this approach ~~[[we]]~~ all related antibody sequences available can be considered. The most fit residue for that position carries a higher ~~value~~~~(e.g.,~~ value (e.g., Koshi et al., 2001, Pac Symp Biocomput 191-202; O. Soyer, M.W. Dimmic, R.R. Neubig, and R.A. Goldstein; Pacific Symposium on Biocomputing 7:625-636 (2002). Sites can be grouped into site-classes or treated independently. Sites and site classes most fit to change based on the substitution rate and the substitutions most favorable based on the fitness can be selected (Fig. 3, Rule 2a). In some embodiments, these values of fitness can then be used directly as a score, as outlined above and in Equation (1) or Equation (2). In some embodiments all sites with a score above a certain threshold value can be selected. For example, a cutoff (threshold) of 0.0 can be chosen (when the normalization of scores sets the wild type residue found in the reference to be 0.0. In some embodiments, all sites with a score below a certain threshold value can be eliminated. Threshold values of 0.0 or below can be eliminated, thereby only including amino changes that have a higher fitness value than the reference wild type amino acid found in that position. In some embodiments, the sites most tolerant to change can be selected by ranking the sites in order of these scores. For example, the most highly scoring site can be selected, or the 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 50, 60, 70, 80, 90 or 100 most highly scoring sites may be selected. In some embodiments, the sites least tolerant to change can be eliminated by ranking the sites in order of these scores. For example, the least highly scoring site can be eliminated, or the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 500, 600, 700, 800, 900 or 1000 least highly scoring sites can be eliminated.

Please amend the paragraph beginning on page 33, line 29, of the specification with the following amended paragraph:

#### 5.1.5 Rules Based on Substitutions ~~[[From]]~~ from Related Antibody Structures

Please amend the paragraph beginning on page 33, line 30, of the specification with the following amended paragraph:

The structures of many antibodies and their variants are also available in the RCSB protein data bank ((2002) *Acta Cryst. D* **58** (6:1), pp. 899-907); and Structural Bioinformatics(2003); P. E. Bourne and H. Weissig, Hoboken, NJ, John Wiley & Sons, Inc. pp. 181-198. The availability of structures can help identify amino acid changes that affect protein function. One way in which they can be used to do so is to avoid changes to the antibody of interest that will not be structurally tolerated by the antibody. Changes computed *in-silico* using energy functions and force fields correlate with experimentally measured free energy changes in the stabilities of proteins. See, for example, Privalov *et al.*, 1988, *Adv Protein Chem* 39: 191-234; Lee, 1993, *Protein Sci* 2: 733-8; Freire, 2001, *Methods Mol Biol* 168: 37-68; and Guerois *et al.*, 2002, *J Mol Biol* 320: 369-87). Therefore, candidate amino acid changes can be modeled into the structure(s) computationally and changes in the free energy computed. These computationally calculated changes in free energies resulting from the substitutions can then be used directly as a score, as outlined above and in Equation (1) or Equation (2). Alternatively, all changes can be selected that increase the free energy of the antibody by less than a certain value. For example, all changes that would increase the free energy by less than 1kCal/mol can be selected, all changes that would increase the free energy by less than 1.5 kCal/mol can be selected, all changes that would increase the free energy by less than 2kCal/mol can be selected, or all changes that would increase the free energy by less than 2.5kCal/mol can be selected. In some embodiments, all changes can be eliminated that increase the free energy of the antibody by more than a certain value. For example, all changes that would increase the free energy by more than ~~1kCal/mol~~ 1 kCal/mol can be eliminated, all changes that would increase the free energy by more than 1.5 kCal/mol can be eliminated, all changes that would increase the free energy by more than ~~2kCal/mol~~ 2 kCal/mol can be eliminated, all changes that would increase the free energy by more than ~~2.5kCal/mol~~ 2.5 kCal/mol can be eliminated. In some embodiments, the best tolerated substitutions can be selected by ranking the sites in order of the predicted increase in free energy. For example, the substitution with the lowest increase in free energy can be selected, or the 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,



36, 37, 38, 39, 40, 50, 60, 70, 80, 90 or 100 ~~substitution~~ substitutions with the lowest increase in free energy may be selected. In some embodiments, the substitutions with the greatest increases in free energy can be eliminated by ranking the sites in order of these scores. For example, the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 12000, 14000, 16000, 18000 or 20000 substitutions with the greatest increases in free energy can be eliminated (Fig. 3, Rule 1b). ~~{CHANGE FIG.}~~

Please amend the paragraph beginning on page 50, line 19, of the specification with the following amended paragraph:

Once substitutions have been selected using expert system 100 (Fig. 2, step 04), and variants have been designed, synthesized and tested for one or more activity or function, it is desirable to use the sequence and activity information from the designed antibody variant set to assess the contributions of substitutions to the one or more antibody activity or function. This process is represented as step 05 in Fig. 2. Assessment of the contributions of substitutions to one or more antibody function can be performed by deriving a sequence-activity relationship. Such a relationship can be expressed very generally, for example as shown in Equation 3

$$(Eq\ 3)\ Y = f(x_1, x_2, \dots, x_i)$$

where,

Y is a quantitative measure of a property of the antibody (e.g., activity),

$x_i$  is a descriptor of a substitution, a combination of substitutions, or a component of one or more substitutions in the sequence of the antibody, and

$f()$  is a mathematical function that can take several forms.

Please amend the paragraph beginning on page 51, line 26, of the specification with the following amended paragraph:

In equation 3, the functional form  $f()$  correlates descriptors of an antibody sequence ( $x_i$ ) to its activity. In a simple embodiment of the invention, the function  $f$  can be a linear combination of  $x_i$ :

$$(Eq. 5) Y = w_1x_1 + w_2x_2 + \dots + w_ix_i$$

where  $w_i$  is a weight (or coefficients of  $x_i$ ).

Please amend the paragraph beginning on page 52, line 30, of the specification with the following amended paragraph:

In some embodiments, modeling techniques are used to derive sequence-activity relationships. Such modeling techniques include linear and non-linear approaches. Linear and non-linear approaches are differentiated from each other based on the algebraic relationships used between variables and responses in such approaches. In the system being modeled, the input data (*e.g.*, variables that serve as descriptors of the antibody sequence), in turn, can be linearly related to the variables provided or non-linear combinations of the variables. It is therefore possible to perform different combinations of models and data-types: linear input variables can be incorporated into a linear model, non-linear input variables can be incorporated into a linear model and non-linear variables can be incorporated into a non-linear ~~models~~ model.

Please amend the paragraph beginning on page 53, line 7, of the specification with the following amended paragraph:

Many functional forms of  $f()$  (Eqn. 3) can be used and the functional form can be combined using weights defined in the knowledge base 108 for analysis. For example, Function  $f()$  can assume non-linear form. An example of non-linear functional form is:

$$Y = w_{12} * x_1 * x_2 + w_{123} * x_1 * x_3 + \dots + w_{nn} * x_n * x_n$$

Please amend the paragraph beginning on page 53, line 19, of the specification with the following amended paragraph:

The data describing variants of the initial antibody can be represented in many forms. In some embodiments, all or a portion of the data is represented in a binary format. For example, representing the presence or absence of a specified residue at a particular position by a “1” or a “0” constitutes a linear binary variable. In another example, representing the presence of a specified residue at one position AND a second specified residue at a second position by a “1” constitutes a non-linear binary variable. In some embodiments, all or a portion of the data is represented as Boolean operators. In some embodiments, all or a portion of the data is represented as principal component descriptors derived from a set of properties. See, for example, Sandberg *et al.*, 1998, J Med Chem. 41, 2481-91. Antibody input sequence data can also use descriptors based on comparison with a sequence profile (*e.g.*, a hidden Markov model, or principal component analysis of a set of sequences). For example in Fig. 9, PC1 and PC2 values of the sequences can be used as descriptors for the sequences in that figure ~~[[Fig.]]~~. In addition, any number of principle components can be used as descriptors. See, for example, Casari *et al.*, 1995, Nat Struct Biol. 2:171-8; and Gogos *et al.*, 2000, Proteins 40:98-105.

Please amend the paragraph beginning on page 56, line 6, of the specification with the following amended paragraph:

The various modeling techniques and algorithms described herein can be adapted to derive relationships between one or more desired properties or functions of an antibody and therefore to make multiple predictions from the same model. Modeling techniques that have been adapted to derive sequence-activity relationships for antibodies are within the scope of the present invention. Some of these methods derive linear relationships (for example partial least squares projection to latent structures) and others derive non-linear relationships (for example neural networks). Algorithms that are specialized for mining associations in the data are also useful for designing sequences to be used in the next iteration of sequence space exploration. These modeling techniques can robustly deal with experimental noise in the activity measured

for each variant. Often experiments are performed in replicates and for each variant there will be multiple measurements of the same activity. These multiple measurements (replicate values) can be averaged and treated as a single number for every variant while modeling the sequence-activity relationship. The average can be a simple mean or another form of an average such as a geometric or a harmonic mean. In the case of multiple measurements, outliers can be eliminated. In addition, the error estimation for a model derived using any algorithm can incorporate the multiple measurements through calculating the standard deviation of the measurement and comparing the predicted activity from the model with the average and estimate the confidence interval within which the prediction lies. Weights for observations to be used in models can also be derived from the accuracy of measurement, for example, through estimating standard deviation and confidence intervals. This procedure can put less emphasis on variants whose measurements are not accurate. Alternatively, ~~theses~~ these replicate values can be treated independently. This will result in duplicating the sequences in the dataset. For example, if sequence variant  $i$  represented by descriptor values  $\{x_j\}^{i1}$  has been measured in ~~triplicates~~ triplicates  $(Y_{i1}, Y_{i2}, Y_{i3})$ , the training set for modeling will include descriptor value  $\{x_j\}^{i2}$  with activity  $Y_{i2}$  and  $\{x_j\}^{i3}$  with activity  $Y_{i3}$  in addition to  $\{x_j\}^{i1}$  with activity  $Y_{i1}$ , where  $\{x_j\}^{i1} = \{x_j\}^{i2} = \{x_j\}^{i3}$ .

Please amend the paragraph beginning on page 57, line 3, of the specification with the following amended paragraph:

*Step 302.* Relevant descriptors of the monomeric variables are identified. These descriptors can convey physico-chemical properties relevant to the interaction between biomolecules or classify the monomers (residues) as discreet entities represented in binary form as described earlier. The former is preferred for residue positions in the antibody sequence where the number of different amino acid substitutions is four or more or where the variables can assume one of four possible values for those positions and the physico-chemical properties values are well distributed (*e.g.*) different from each other. The latter is preferred for positions that have four or less possible values for the relevant variable, and/or the values are clustered (*e.g.*) are not very different from each other. To create non-linear variables, new variables are formed that are a combination of monomeric variables. For example, consider two variants AGWRY and AKYRY. The linear binary form of the variable (descriptor) for

position 2 assumes a value of “1” if G is present at that position and “0” if it is absent. Alternatively, a non-linear variable can be created in addition to the linear variables describing each position. In the above example, a new non-linear variable representing position “2” and “3” can assume four values in numeric form. In one form, the variable can assume a value of 11 for “GW”, 10 for “GY”, 01 for “KW” and 00 for “KY”. In other representations of binary non-linear variable, four variables can describe position 2 and 3, where variable one assumes a value of “1” if the sequence at position 2 and 3 is “GW” and “0” otherwise and the second variable takes the values of “1” or “0” if the sequence is “GY” or otherwise and so on.

Please amend the paragraph beginning on page 58, line 4, of the specification with the following amended paragraph:

*Step 304.* In step 304 the parameters for the functional form of the sequence-activity relationship are optimized to obtain a model by minimizing the difference between the predicted values and real values of the activity of the antibody. Such optimization adjusts the individual weights for each of the descriptors identified in preceding steps using a refinement algorithm such as least squares regression techniques. Other methods that use alternative loss functions for minimization can be used to analyze any particular dataset. For example, in some antibody sequence-activity data sets, the activities may not be distributed evenly throughout the measured range. This will skew the model towards data points in the activity space that are clustered. This can be disadvantageous because datasets often contain more data for antibody variants with low levels of activity, so the model or map will be biased towards accuracy for these antibodies that are of lower interest. This skewed distribution can be compensated for by modeling using a probability factor or a cost function based on expert knowledge. This function can be modeled for the activity value or can be used to assign weights to data points based on their activity. As an example, for a set of activities in the range of 0 to 10, transforming the data with a sigmoidal function centered at five will give more weight to sequences with activity above five. Such a function can optionally also be altered with subsequent iterations, thereby focusing the modeling on the part of the dataset with the most desired functional characteristics. This approach can also be coupled with exploring techniques like a Tabu search, where undesired space is explored with lower probabilities.

Please amend the paragraph beginning on page 59, line 14, of the specification with the following amended paragraph:

*Step 308.* In step 308 the coefficients (parameters) of the model(s) are deconvoluted to see which amino acid substitutions (variables/descriptors of the variants) influence the activity of the antibody. It can be important to identify which descriptors of the antibody are important for the activity of interest. Some of the techniques, such as partial least squares regression (SIMPLS) that uses projection to latent structures (compression of data matrix into orthogonal factors) may be good at directly addressing this point because contributions of variables to any particular latent factors can be directly calculated. See, for example, Bucht *et al.*, 1999, *Biochim Biophys Acta*. 1431:471-82; and Norinder *et al.*, 1997, *J Pept Res* 49:155-62. Other methods such as neural networks can learn from the data very well and make predictions about the activity of entire antibodies, but it may be difficult to extract information, such as individual contributing features of the antibody from the model. Modeling techniques/methods that directly correlate the amino acid variations to the activity are preferred because we can derive the sequence-activity map (relationship) to construct new variants not in dataset that have preferentially higher activities. These methods can be adapted to provide a direct answer and output in desired forms.

Please amend the paragraph beginning on page 62, line 14, of the specification with the following amended paragraph:

It will be appreciated by one skilled in the art that each different method for deriving relationships between antibody sequences and activities can differ in the precise values of their outputs. In some embodiments of the invention it is therefore desirable to combine the outputs from two or more such methods for subsequent uses. This corresponds to step 06 in Fig. 2. There are a variety of ways in which such outputs can be combined. In some embodiments, each output can be independently applied to the subsequent design of antibody variants (Fig. 2, step 07) or the modification of parameters or weights used by expert system 100 for the selection of substitutions (Fig. 2 step 02) or the design of antibody variant sets (Fig. 2 step 03). In some embodiments, average values (or some other mathematical function of two or more

values derived by two or more sequence-activity models) can be calculated for the regression coefficient, weight or other value describing the relative or absolute contribution of each substitution or combination of substitutions to one or more activity of the antibody (*e.g.*, as defined in Equation 4 below). In some embodiments, the standard deviation, variance or other measure of the confidence with which the value describing the contribution of the substitution or combination of substitutions to one or more activity of the antibody can be assigned (*e.g.*, as defined in Equation 4 below). In some embodiments, the rank order of preferred substitutions is used to combine the methods. In some embodiments, the additive (linear variables) and non-additive components (non-linear variables) of each substitution or combination of substitutions is combined:

$$(Eq. 6) V_{ix} = f(M_1(i_x), M_2(i_x), \dots, M_j(i_x))$$

where,

$V_{ix}$  is a combined measure of one of the descriptors measuring the performance of an antibody in which monomer  $x$  is substituted at position  $i$ ;

$M_j(i_x)$  is a measure of one of descriptors measuring the performance of an antibody in which monomer  $x$  is substituted at position  $i$ , determined by sequence-activity correlating method  $j$  ( $M_j(i_x)$  is the contribution of  $i_x$  as determined by Model  $j$ ); **and**

$f()$  is some mathematical function.

Please amend the paragraph beginning on page 122, line 1, of the specification with the following amended paragraph:

Aspects of the present invention can be implemented as a computer program product that comprises a computer program mechanism embedded in a computer readable storage medium. For instance, the computer program product could contain the program modules and/or data structures shown in Fig. [[2]] 1. These program modules may be stored on a CD-ROM, magnetic disk storage product, digital video disk (DVD) or any other computer readable data or program storage product. The software modules in the computer program

product may also be distributed electronically, via the Internet or otherwise, by transmission of a computer data signal (in which the software modules are embedded) on a carrier wave.